DOCUMENT RESUME

ED 278 984 CS 008 725

AUTHOR Pearson, P. David; Dunning, David
TITLE The Impact of Assessment on Reading Instruction.
PUB DATE 85
NOTE 13p.
PUB TYPE Guides - Classroom Use - Guides (For Teachers) (052) -- Journal Articles (080)
JOURNAL CIT Illinois Reading Council Journal; v13 n2 p19-29 Fall 1985

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Class Activities; Curriculum Problems; Elementary Secondary Education; Evaluation Methods; *Mastery Learning; Mastery Tests; Reading Achievement; Reading Diagnosis; *Reading Instruction; Reading Skills; *Reading Tests; Skill Development; *Standardized Tests; Student Evaluation; Teaching Methods; Test Construction; Test Format; *Testing Programs; Test Interpretation; Test Reliability; Test Use
IDENTIFIERS *Teaching to the Test

ABSTRACT

Schemes for assessing reading achievement with specific tests have been in use since early in this century. The two driving forces behind the testing movement--scientific objectivity and compulsory education--blossomed in the 1920s and 1930s and continued substantially unchanged through most of the 1960s. In the early 1970s, the idea of mastery learning (hold achievement constant and allow instruction to vary) was introduced and developed in the form of highly skill-specific, criterion-referenced tests. Generally, instructional variation consisted of varying the amount of workbook or worksheet practice designed to help students pass mastery skills tests. Since the late 1970s, states have shifted their concern to instruction and program assessment. In the early 1980s, the Report of the Commission on Reading caused an increase in comprehensive state-mandated testing programs (SMTP). The major threat of these programs is curriculum reductionism, which teachers might respond to by (1) not organizing classroom time around "teaching to the test"; (2) learning more about the process for revising SMTP; (3) interpreting SMTP results cautiously because they are generally not an accurate index of an individual's progress; and (4) evaluating children's reading competency using measures that approximate real reading. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED278984

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The Impact of Assessment on Reading Instruction

P. David Pearson

David Dunning

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Evelyn W. Jackson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

CS 008 725

THE IMPACT OF ASSESSMENT ON READING INSTRUCTION

P. David Pearson and David Dunning

Early origins

Most of us fix the responsibility for siring assessment schemes for measuring reading achievement to Edward L. Thorndike around the time of the First World War. And it is probably a fair assignment of paternal responsibility if, by assessment, we mean the development of standardized tests. However the notion of independent and unbiased assessment by trained experts is really much older, dating back, at the least, to medieval times and the "trial by ordeal" framework of the examination for the doctor of philosophy degree in German and English universities (a ritual that survives, pretty much intact, even today).

In fact, one can imagine that Thorndike and his cohorts thought they were providing a great service to students by shifting the authority over "rites of passage" from the hands of a few highly trained and highly insular individual examiners to the purity of an objective and unbiased set of multiple-choice questions on a sheet of paper. One wonders if they could have envisioned, in their most vivid nightmares, the vast array of standardized examinations that we document every half-decade or so in a new edition of the *Buros' Mental Measurement Yearbook*.

If both the form of the standardized test (multiple-choice items) and the quintessential feature of tests, reliability, are legacies from Thorndike's work, it is to another pioneer, Joseph Mayer Rice, that we owe our thanks for the idea of assessing instructional effectiveness with specific tests, such as spelling, reading or handwriting scales.

The two driving forces behind the testing movement that blossomed in the Teens and Twenties were scientific objectivity and compulsory education. The relationship between "science" and testing is transparent; tests were thought to be more objective than personal judgments. The contribution of compulsory education was that it created a need to "monitor" the high rate of failure that our schools suddenly experienced when they no longer had the option of dismissing low achievers from our schools. And surely the development, in 1935, of the IBM 805 test scanner nailed down the place of objective multiple-choice tests by reducing the cost of scoring to 1/10th that of hand scoring.

And so our assessment of reading, as it developed from the Twenties to the Sixties, changed only within a well-established framework—objective, machine-scorable, multiple-choice tests that exhibited high reliability. The issue of validity (the degree to which a test measures the trait it purports to measure) was, of course, considered, but usually it was addressed by establishing face validity (does

this look like what we define reading to be) or concurrent validity (how well does this new test correlate with well established, and usually highly regarded, tests of reading).

There was some important work done within this framework during the 1920-1965 period. Frederick Davis and others confronted the issue of skill uniqueness. What Davis wanted to know was whether or not the 7 or 8 or 10 or 20 different subskills of reading included in a reading test were really psychometrically unique (distinct from one another). So he used the emerging technique of factor analysis to test uniqueness, finding that among comprehension skills, the set usually reduced to a few distinguishable clusters such as a word (vocabulary) factor, a gist (main point) factor, and a reasoning (inference) factor. At the end of this period, researchers like Tuinmann and Oaster uncovered the startling finding that many reading test items could be answered without reading the text accompanying the items.

Mastery learning

The next watershed development in reading tests began in the late Sixties and reached full momentum in the mid-Seventies. It was the development of detailed and highly skill-specific criterion-referenced tests, tests that dissected the reading process into its most atomistic elements and provided an index of performance on each element. The movement was driven by the notion of mastery learning, introduced by John Carroll in the early Sixties and developed fully by the work of Benjamin Bloom at the end of the decade. The basic point of departure for mastery learning was the criticism that what we had done in our schools was to hold instruction constant across students of varying aptitudes (teach everyone *equally*) and allow achievement to vary. Instead, what we should do is to hold achievement (or at least some minimal, i.e., mastery, level of achievement) constant and allow instruction to vary.

Buttressed by the egalitarian appeal of such an approach, many scholars and publishers devoted enormous energy, from 1968 to 1980, to constructing tests emanating from this mastery model. But in adopting the mastery model, these test developers incidentally embraced another guideline for constructing their tests: virtually all of them embraced a sub-skills view of the process of learning to read. Hence, instead of applying the mastery notion to wholistic or integrated reading behaviors, they applied it to all of the *enabling* skills that basal reading series usually teach across grade levels in their attempt to build an expert reader. So instead of building mastery tests that measured the ability to understand, say, fourth- or fifth- or tenth-grade material of some genre, they built tests of vowel di-

graphs, consonant blends, main idea, sequence, cause-effect, map reading, skimming for details, etc. The net result of some of these mastery learning assessment programs was to test from 50 to 300 separate skills from grade 1 through grade 8.

Along with mastery learning and a vast array of discrete subskills came a shift from norm-referenced to criterion-referenced standards for evaluating individual performance. From 1915 to 1970, virtually all widely used published reading tests (dubbed "standardized" tests by we know not whom) provided users with a "norm" to which any individual's (or any group's) performance could be compared in order to make a judgment about how well or poorly that individual (or group) was performing. The norm was almost always the actual or extrapolated (i.e., inferred) mean of some identifiable "norm group." So, for example, when a teacher learned that Suzy Jones, an eighth grade student in his class, achieved a grade-norm score of 6.8, he could infer that she was performing about as well as the average sixth grade student who took the test during the eighth month of school. Or if he learned that Sam Hernbill scored at the 47th percentile, he could infer that, compared to a national norm group of eighth grade students, Sam did as well or better than 47% of those students. But with mastery learning and criterion-referenced testing came an alternative benchmark for evaluating performance. Instead of comparing an individual's score with the mean score of some group, advocates of the new approach suggested comparing it with some pre-determined absolute standard of mastery, say 80% correct on the items for a given subtest. This kind of standard is, of course, consistent with the guiding principle of mastery learning: hold achievement constant and allow instruction to vary!

But this was only one part of the mastery learning system. These tests allowed teachers to hold achievement constant. But how were they to accomplish the feat of allowing instruction to vary? In nearly all the mastery learning systems with which we are familiar, there has been but a single operational definition of instructional variation: varying amounts of workbook or worksheet practice. So it was in the first half of the decade of the Seventies that we saw classroom after classroom in which the following drama was enacted on a daily basis:

A group of students took a mastery test on a battery of skills deemed appropriate for their grade level.

The teacher (or a machine) scored the test and provided a list of which kids had failed to pass which skills. From these lists the teacher created specific skill groups.

Students were given a list of a set of workbook pages and ditto sheets that they were to complete on that skill. The students went to the set of bins along the wall, selected the appropriate worksheets, returned to their desks, completed the worksheets, and turned the set into the teacher. The teacher, during this time, spent most of her energy engaged in the laborious task of grading the worksheets that the students were completing.

When an individual or a subset of students appeared to be ready to retake a particular mastery test or a battery of tests, they did so. Those who passed went on to conquer new tests and new skills; those who failed a second time were assigned new sets of worksheets to complete in the hope that the additional practice would be just the right antidote for their repeated failure.

To those who would judge our scenario to be an exaggerated caricature we respond that those classrooms were there; we saw them and continue to see them. The great irony of these mastery learning systems, when variable instruction is limited to varying amounts of practice, is that they result in a "the rich get richer and the poor get poorer" outcome. This is an inevitable consequence of repeated practice on any cognitive task: people get better at what they practice most, and if what they practice is how to perform a task inappropriately, then they get better at inappropriate performance.

Well, it did not take long for the basal reading companies to see the wisdom of the mastery learning measures that independent companies had already been publishing. They soon learned that schools with these mastery learning testing systems wanted to know at what points and on what workbook and ditto master pages the skills tested in the mastery learning system were "taught" in the basal series. So they responded in two ways. First, they provided what we have come to call *correlations*, lists that specify the levels and pages on which the skills assessed in the school's testing program are "taught." Second, they began to play the mastery testing game themselves. Prior to this period, basal reading series had check-up tests at the end of units within a level in a series, but neither the publishers nor the teachers made much of a fuss about them; they were there if you wanted to make use of them, but no one seemed to want to use them to make any grave decisions of any consequence to an individual child. But by the mid-Seventies, every basal publisher had created its own set of mastery tests to compete with the series-independent versions. Furthermore, the companies began to bring out supplemental workbooks, booster packs of ditto masters, and practice pads so that teachers would have something to give to the children who did not do so well on the end-of-unit mastery tests. And in the "toughest" of school districts, students who did not pass the end-of-unit tests had to

repeat all the stories and all the worksheets in the unit before they could take the mastery test again.

We have written this section as if these activities existed in the past. Alas! They are all too real in the schools of 1985. The tests still exist, the worksheets still exist, and the evidence mounts daily that virtually all that goes on in the name of "instruction" for even the most complex of reading skills is practice operationalized in the form of worksheets. And desperate students still fail daily despite our rhetoric about mastery.

But our history is not yet complete. There are other characters and other movements that we must examine in order to account for where we are today.

States get into the act

According to George Madaus, our angry concerns about the evils of criterion-referenced testing systems spawned by mastery learning ideals is of little consequence. The manner in which testing informs decision-making within the classroom is no longer the issue anyway. The power for making key policy setting decisions regarding the connection between instruction and assessment does not rest in the hands of classroom teachers; nor is it within the grasp of local school district administrators. Rather, it is the states that run the show nowadays, and they have been doing so increasingly since the early Seventies when the state-level emphasis on guidance testing began to shift to a concern for instruction and program assessment, largely in response to the need for state agencies to monitor the education programs funded by federal money.

Although there are probably no typical state-level testing programs, nearly all the states have a minimum competency test or assessment testing program or requirement, and nearly all of the state-mandated tests cover skills in both reading and math. In addition, most of the testing programs have been developed or contracted out by the states themselves with an eye towards approximating an "average" instructional program for the state. So, in addition to being home-grown in the majority of circumstances, most of the state assessment tests appear to be criterion-referenced (or what many call objectives-referenced in deference to the necessary step of creating objectives in advance of writing the test items) in design as well. Given these last two conditions, it is theoretically possible for states to bring pressure on local school districts to bring their local curricula into line with a state standard.

State assessment information is already being used by some states, e.g. Delaware and California, to remedy perceived statewide problem

areas in specific subject skills. In California, for instance, statewide education officials, in examining state-level assessment results, concluded that sixth-grade achievement in math computation was down. As a result they revised the mathematics curriculum framework for the state, and soon the textbook guidelines for the state reflected this change in focus. Subsequent to this change it was noticed that scores in math problem solving had dropped off and a similar process was initiated all over again.

Many state education officers, legislators and state education board members view the state assessment testing program as a "tool for accomplishing educational goals" for their state. This tool can be a powerful force for affecting instructional focus and curriculum and can have a constricting effect when coupled with the effect which in Europe is known as "the tradition of past exams," that is, the influence which previous exams have on the current curriculum. So given a predictable and specified content on statewide exams, it should not be surprising to find widespread "teaching to the test," especially if students' educational futures and/or teachers' careers are "on the line" (or at the very least, students and teacher *think* they are "on the line").

The "Commission" reports

The last two years have witnessed the entry of a new player in this drama—the "Commission Report." At the state level the response to the recent flurry of national commission reports bemoaning the condition of our educational system seems to have been a "rising tide" of increasingly comprehensive state-mandated programs for student testing. Often included as part of overall state aid and reform packages, these testing programs appear to be legislators' way of ensuring that increased accountability accompanies the increased funding being requested for educational programs.

In Illinois, for example, a nearly $400 million proposed school aid plan, which includes overall provisions such as general aid to local schools as well as more specific ones such as state funding for optional kindergartens and preschools, also includes stipulations for student testing in reading, math and language arts at grade levels 3, 6, 8, and 10, with special remedial programs for students whose scores fall more than one year below grade level (The News Gazette, 1985). Significantly, "learning outcomes" — a general outline of what students should know when they complete a given level of schooling — are being proposed to replace course requirements. It seems inevitable that the evaluation of these learning outcomes will become a "natural" part of the state-wide testing program.

There are many problems with externally imposed testing programs, not the least of which is the apparent arbitrariness of any given cut-off points for those students falling just below them.

Teachers also need to be cautious about the increasing competitive climate among teachers which may be fostered as a result of the increased importance attached to the results of a state testing program. By contrast, one of the findings of school effectiveness research is that ". . .the most effective schools — where student learnings are greatest — do not isolate teachers but instead encourage professional dialogue and collaboration" (Rosenholtz, 1985, p. 351). An excessive concern with classroom-by-classroom results for state testing programs, then, could seriously disrupt the kind of collegial atmosphere which seems to contribute to making schools effective.

Where we are now: How to make assessment work

Perhaps the major threat which state-mandated testing programs pose for our educational system, however, lies in the area of curriculum and the tendency for any externally imposed testing program to narrow the focus of classroom instruction to that collection of skills which is covered on the test (what we have called the tradition of past exams).

There are some educators, nonetheless, who approve of this narrowing of curricular focus and argue that measurement-driven instruction results in higher achievement in the subject area tested in. Popham and others (1985) describe the criterion-referenced testing programs of Texas, Detroit, South Carolina and Maryland and note approvingly that although they differ in many ways, the programs share the one common feature that measurement was perceived as a "catalyst to improve instruction" (Popham et al., 1985, p. 628).

Unfortunately, measurement also appears to impose a catalyst for providing constraints on instruction, and teaching to the test seems to be not only a common but an encouraged practice. In Maryland, for example, "diagnostic quizzes" are regularly administered to the students as part of the preparation procedures for taking the tests. In addition, state-sponsored inservice programs have supported the production of large amounts of "high-quality, on-target" instructional materials for use in the classroom (Popham et al., 1985). In offering an explanation for the climb in Maryland's test scores, the chief of educational program assessment for the state notes that "much of this consistent improvement appears to stem from the increasing familiarity of Maryland educators with the test specification for the skills being assessed" (Popham et al., 1985, p. 634). For Maryland educators, then, knowing what is on the test, teaching to the test and

using instructional materials geared to the test's content are all state sanctioned and state encouraged approaches to teaching.

Narrowing instructional focus for the purpose of improving performance on a test is not necessarily an ill-conceived strategy, however. If the test represents a valid measure of the skill under consideration, then teaching to the test makes a lot of sense. For example, in the area of aviation, if the exam for determining a candidate's fitness for becoming a commercial pilot were held in a flight simulator (which we assume that it is) then we might find little reason for altering our future travel habits to include fewer flights. But if we suddenly learned that the only federally required test of, say, a commercial aviation pilot's ability to fly a plane was a one hundred item multiple-choice exam, then perhaps our next long distance trip might be by AMTRAK!

The Commission on Reading puts it well: "Holding a reading teacher accountable for scores on a test of say dividing words into syllables is like holding a basketball coach accountable for the percentage of shots players make during the pre-game warm up" (Anderson et al., 1984, p. 99). To the extent that state-mandated reading tests assess proficiency in "real-world" reading, that is, the type of reading which students and adults encounter beyond the classroom walls, then they are valid indicators of true reading ability. Unfortunately, tests of atomistic reading skills in a multiple-choice format are hardly representative of the types of reading demands facing students either in their further schooling or in life beyond the classroom.

A couple of comments are in order here about how we, as classroom teachers, might consider responding to the specific threat of curricular reductionism posed by state-mandated tests.

First, as teachers, we should be careful about how we spend our time in the classroom and resist organizing our instruction primarily around the tests. True, state-mandated assessment is likely to be around for quite awhile. And in some states, such as Maryland, instructional goals are being more tightly focused around the content of the tests. But as teachers our first responsibility should continue to be to the students and only secondarily to the state testing boards. And whether or not we choose to spend class time preparing for a state test, we should ensure that students increase their proficiency in competencies which, while probably being among the more valid indicators of reading proficiency, will probably not be found on state-mandated tests. These sort of basic reading proficiencies, as noted by the Commission on Reading, would include the ability to ". . .read aloud unfamiliar but grade-appropriate material with acceptable fluency; write satisfactory summaries of unfamiliar selections from grade-appropriate social studies and science textbooks; explain the

plots and motivations of characters in unfamiliar, grade-appropriate fiction; read extensively from books, magazines, and newspapers during leisure time" (Anderson et al., 1984, p. 99).

A second consideration that teachers might keep in mind is to learn more about the process for revising state-mandated tests (if there is one for their state), and if possible to lend encouragement to test board members for including items which are the closest possible approximation to the type of functional reading which the Commission on Reading and all sensible educators would encourage.

Living with assessment

While there may be disagreement about the nature of the impact of state-mandated testing programs on classroom instruction, few educators would disagree that feedback regarding student's performance is both a natural and a necessary component of good instruction. And while large group testing programs do have limitations in their ability to provide an accurate index of individual students' reading ability, still, testing results can legitimately be used as one source of information for making diagnostic decisions about instruction in the classroom. Specific decisions about individual students can draw on multiple sources of information—teachers' informal observations in reading groups, informal reading inventories as well as students' performance on standardized tests.

Criterion-referenced tests are often cited as being an effective diagnostic instrument providing useful information on an individual student's performance. Teachers should exercise caution in the interpretation of the results, however, because, while criterion-referenced tests are reasonable indicators of the local effects of class instruction, they are less accurate as an index of an individual's progress. The reason for this is that many of them don't contain enough test items within each tested subskill to make the overall results reliable. In other words, a student's performance on the various subtests is likely to shift considerably from one administration of the test to the next. Perhaps such tests would be better used to evaluate *instruction* rather than *learning*.

Perhaps a more reliable measure of a student's reading ability might be obtained by taking repeated samples of his/her reading performance in one-on-one situations. In contrast to the kind of oral reading which can occur in round robin reading situations where the bad habits of the poorer readers are broadcast to everyone's embarrassment, one-on-one reading provides both the student and the teacher the opportunity as well as the privacy to focus on the individual student's reading needs. The quality of a teacher's observations made under these conditions is likely to have greater diagnostic

value than either those made under the generally more harrowing circumstances of group round robin reading or those made while reviewing the printouts from a standardized testing program. Teachers should have more confidence then in their ability to make judgments about children's reading proficiency when based on repeated exposures to student's reading performance in one-on-one situations.

As teachers, then, we need to make sure that our ultimate evaluation of children's reading competency approximates real reading as closely as possible. In other words, our judgments about students' reading competencies should be based not on how well they complete various worksheets dealing with atomistic reading skills, but rather we should be focusing on how well they are putting their reading ability to use in meaningful and functional ways — for example, being able to read the signs in their neighborhood, reading age-appropriate fiction for pleasure, and being able to look up information in reference books.

So even in schools where state-mandated testing is a way of life and administrative supports for encouraging teaching to the test behaviors are firmly in place, as long as teachers can ensure that their students are adequately prepared for the test, what teachers do behind their closed doors is still largely up to them. A fourth grade teacher, for instance, whose students consistently perform well on standardized testing should not feel intimidated if she wants to conduct a literature-based reading program for her students. Rather, she should feel confident that she is providing a meaningful and appropriate reading curriculum.

As teachers, we must resist being sidetracked from important educational goals, even if these sidetracking influences come in the form of officially mandated testing programs. If the tests indeed are "broad-gauged measures that reflect the ultimate goals of instruction" (Anderson et al., 1984, p. 100), then well and good. Perhaps we should not be too reluctant then to teach to the test. But such is often not the case. As classroom teachers we are the final arbiters of our students' education, operating as we do on the front lines of instruction; we should take the opportunity provided to us to ensure that the content of that instruction is the best that we can devise. We owe our students no less.

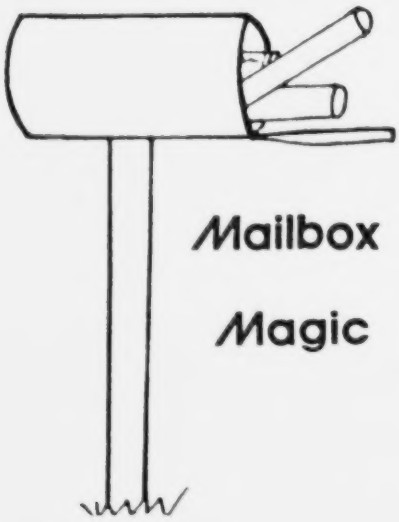
REFERENCES

Anderson, R.C., Hiebert, E.H., Scott, J.A., & Wilkinson, I.A.G. (1985) *Becoming a nation of readers: The report of the commission on reading.* National Institute of Education.

Popham, W.J., Cruse, K.L., Rankin, S.C., Sandifer, P.D., & Williams, P.L. (1985) Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66*, 628-634.

Rosenholtz, S.J. (1985) Political myths about education reform: Lessons from research on teaching. *Phi Delta Kappan, 66*, 349-355.

"Highlights of School Funding Plan." *News Gazette*, (July 3, 1985), p. 2.



Selected magazines for children recommended by members and friends of the ILLINOIS READING COUNCIL.

Printed black on blue, this magazine list for preschoolers through junior high may be ordered from Roxanne Henkin, 223 West Country, Bartlett, IL 60103.

001 for $0.05 + stamped envelope
025 for $1.00 + $0.50 shipping
100 for $3.00 + $2.00 shipping

END

U.S. DEPT. OF EDUCATION

OFFICE OF EDUCATIONAL RESEARCH AND IMPROVEMENT (OERI)

ERIC

DATE FILMED

JULY 14 1987